Thought Experiments in the Age of Frontier AI: Capability, Gatekeeping, and the Future of Scientific Practice

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

December 9, 2025

Frontier-level AI systems have quietly made a new kind of scientific practice possible. For any domain that can be expressed in language, they can now serve as fast, articulate partners in constructing and probing thought experiments: exploring counterfactual worlds, staging multi-perspective debates, stress-testing hypotheses, and mapping edge cases that human collaborators might never reach in time. Crucially, in this "thought-experiment regime" there is no direct wet-lab risk, no immediate patient harm, and no autonomous deployment. The human scientist still sits at the decision boundary, responsible for what is believed and what is done next. Yet this possibility is far from normalized. Younger scientists report a strong informal chill: it is acceptable to use AI to polish prose or fix code, but risky to admit using it to generate or refine core ideas. The pattern is sociological as much as technical. Established hierarchies are defending a cognitive monopoly, a narrative about how real science is done, and a labor structure built for a pre-AI world. This paper analyzes that tension and sketches what an honest, robust practice of AI-assisted thought experiments could look like, including concrete norms and templates that younger scientists might adopt without surrendering either rigor or responsibility.

Keywords: frontier AI, thought experiments, scientific practice, centaur intelligence, epistemic gatekeeping, generational dynamics, scientific norms, peer review, authorship, AI adoption, institutional inertia, research ethics, methodological transparency.

A collaboration with GPT-5.1-Thinking-Extended. 41 pages. CC4.0.

Outline

1. Introduction
1.1 Frontier AI and the quiet expansion of thought-experiment space
1.2 Why the controversy is sociological, not technical, in the safe regime
1.3 Scope, methods, and structure of this paper

2. Frontier AI as a Thought-Experiment Instrument
2.1 From chat interface to conceptual laboratory
2.2 Types of AI-assisted thought experiments across disciplines
2.3 The “safe region”: ideas, not actions or deployments
2.4 Human responsibility at the acceptance boundary

3. How Scientists Are Actually Using AI Today
3.1 Surface adoption: editing, summarization, and boilerplate code
3.2 Hidden adoption: private notebooks and unacknowledged ideation
3.3 Differences across fields, career stages, and geographies
3.4 The emerging double life: AI-rich practice, AI-poor narratives

4. Gatekeeping and Cognitive Monopoly Protection
4.1 Senior cognitive advantages and how AI erodes them
4.2 Narrative control over what counts as “real thinking”
4.3 Labor, credential, and status preservation in AI discourse
4.4 Peer and peer-review pressures on younger scientists

5. Legitimate Concerns Versus Protective Rationalizations
5.1 Epistemic slippage and the problem of hallucinated structure
5.2 Fears of shallow dependence and lost training value
5.3 Accountability and authorship when AI enters idea pipelines
5.4 Distinguishing real risks from hierarchy-preserving myths

6. Centaur Practice: Positive Norms for AI-Assisted Thought Experiments
6.1 Defining the centaur: human judgment plus machine breadth
6.2 Principles for legitimate AI use in conceptual work
6.3 Transparency without fetish: describing AI as method, not magic
6.4 Evaluating originality when some scaffolding is machine-generated

7. Practical Templates for Younger Scientists
7.1 A model “AI-assisted thought experiments” methods subsection
7.2 Disclosure patterns that are honest but not self-sabotaging
7.3 Lab-level norms: group charters for centaur use
7.4 Navigating peer review: responding to objections about AI use

8. Implications for Institutions and the Future of Scientific Work
8.1 Rethinking training: teaching AI-literate thought-experiment design
8.2 Granting, hiring, and promotion in a centaur world
8.3 From taboo to infrastructure: specialized tools for conceptual labs
8.4 Long-term trajectories: from lone minds to networked intelligences

9. Conclusion
9.1 What is already possible, and what is being blocked
9.2 The costs of prolonged hypocrisy and under-use
9.3 A constructive path: normalizing AI-assisted thought experiments
9.4 Final reflections on courage, responsibility, and shared intellect

10. References

1. Introduction

The arrival of frontier-level AI systems has quietly altered what it means to think in a scientific context. In any domain that can be articulated in language, a researcher can now sit down with a model and, within minutes, generate and interrogate elaborate thought experiments: counterfactual worlds, multi-perspective debates, adversarial critiques of their own hypotheses, and systematic scans of edge cases that used to take weeks of hallway conversations and draft exchanges. The core novelty is not that AI supplies answers, but that it radically lowers the cost of structured imaginative work.

At the same time, this new practice remains largely unofficial. It is common and uncontroversial to acknowledge AI for proofreading, code refactoring, or formatting help. It is much rarer to see a methods section that states plainly that a model was used to generate alternative explanations, stress-test arguments, or explore families of hypothetical scenarios. The discrepancy is striking: in private, many younger scientists treat AI as a conceptual partner; in public, the story told in papers and grant applications still resembles the pre-AI world.

This paper takes that tension as its object. It argues that within a specific, conceptually rich but operationally safe regime, the main obstacles to AI-assisted thought experiments are not technical or ethical in the narrow sense. They are sociological. Incumbent hierarchies are defending a bundle of privileges: cognitive monopoly, narrative control over what counts as real science, and a labor structure built around older assumptions about how many minds are needed to do serious work. The goal here is not to demonize senior scientists, but to make these dynamics visible and to sketch a more honest, constructive path for integrating AI into the life of inquiry.

1.1 Frontier AI and the quiet expansion of thought-experiment space

Thought experiments have always been a central tool of science and philosophy. Imagined falling elevators, demonically fast sorters of molecules, Schrödinger's cat, hypothetical randomized trials that no ethics board would approve: these constructs sharpen concepts, reveal hidden assumptions, and test the internal coherence of theories. Historically, they have depended on the unaided human imagination, occasionally amplified by informal conversation or mathematical toy models.

Frontier AI systems extend this practice in three important ways.

First, they make it trivial to instantiate multiple perspectives. A single researcher can now ask a model to play advocate, skeptic, and meta-analyst in turn; to argue from the standpoint of different schools or traditions; to reframe a problem as seen by a clinician, a statistician, and a patient. This multi-perspective capacity used to require time-consuming collaboration. Now it can be invoked on demand, iterated quickly, and logged as a trail of reasoning.

Second, they expand the reachable periphery of hypothesis and scenario space. When prompted carefully, a model can generate long lists of alternative explanations, possible confounders, boundary conditions, or failure modes that might not occur to a single human thinker under time pressure. Many of these will be trivial or wrong, but some will be genuinely surprising and useful. The key is not that the model is infallible, but that it functions as a wide-angle lens on possibility space.

Third, they make it easier to systematize variation. Instead of vaguely musing about "what if we changed this assumption," a researcher can ask the model to enumerate, for example, ten parametric changes to a trial design, or five morally relevant modifications of a classic ethical dilemma, and then walk through the consequences of each. The result is not a simulation in the narrow quantitative sense, but a structured exploration of how conclusions depend on the scaffolding of the scenario.

All of this sits upstream of experiment, deployment, or policy. In this mode, AI is not an autonomous actor in the world. It is a configurable generator of hypothetical structures and arguments, whose outputs are then accepted, rejected, or refined by the human scientist. The expansion of thought-experiment space is therefore quiet: it does not show up as a new kind of instrument in the lab, but as a different texture to the reasoning that precedes and surrounds formal work.

1.2 Why the controversy is sociological, not technical, in the safe regime

There are well-founded concerns about using AI in domains where outputs directly touch patients, infrastructures, or security-sensitive operations. Those debates rightly focus on reliability, robustness, and misaligned incentives. In the narrow regime considered here, however, those worries are much less acute. When models are used solely to generate and critique hypothetical constructions, the human scientist remains the gate through which any idea must pass before it becomes an experiment, a protocol, or a claim in the literature. The risk is primarily epistemic, not operational.

Technically, the limitations of frontier AI in this regime are well understood. Models hallucinate, overgeneralize, and sometimes offer plausible but wrong chains of reasoning. They are also capable of self-correction when cross-examined and of flagging their own uncertainty when prompted. In other words, they behave less like an oracle and more like a very fast but uneven collaborator. The remedy is not prohibition, but disciplined use: explicit cross-checking, clear separation between generation and acceptance, and a willingness to treat AI contributions as proposals rather than truths.

If the main technical risk is manageable, why the chill around acknowledging conceptual use? The answer lies in professional incentives and identity. Senior scientists, who largely control

hiring, tenure, and peer review, built their careers in an environment where scarce human cognitive labor was the core resource. Their comparative advantage lay in being able to think more broadly, more deeply, and more articulately than their peers. The emergence of tools that partially level that playing field, especially for younger or less-resourced colleagues, implicitly threatens that advantage.

As a result, disputes that present as concerns about rigor or originality often carry an unspoken subtext: anxiety about erosion of status, authority, and the familiar shape of scientific life. It is easier to declare certain uses of AI illegitimate than to renegotiate authorship norms, redefine what counts as independent thinking, or admit that a junior researcher plus an AI system may sometimes outstrip the ideational throughput of a traditional team. The controversy in the safe regime is therefore less about what the models can or cannot do, and more about who will be allowed to do what with them, under whose narrative.

1.3 Scope, methods, and structure of this paper

This paper focuses on the use of frontier AI for thought experiments in scientific and scholarly work, under three deliberate constraints.

First, it treats only the conceptual regime: the spacing of hypotheses, counterfactual reasoning, design of hypothetical scenarios, exploration of moral and methodological edge cases. It does not address AI-driven automation of data collection, robotic control, clinical decision support, or autonomous experimentation, where the risk landscape is very different.

Second, it brackets questions of model training and alignment except where they directly bear on epistemic reliability in conceptual work. The aim is not to evaluate specific architectures or vendors, but to examine how currently available systems function as instruments for thinking, and how their use is being socially negotiated.

Third, it approaches the topic as a hybrid of normative and descriptive analysis. Descriptively, it draws on observed patterns of AI use in research practice, including the emerging distinction between publicly acceptable and privately common uses. Normatively, it argues for a set of positive norms around what might be called centaur practice: configurations in which human judgment and machine generativity are explicitly combined, with the human remaining accountable for what is endorsed.

The structure of the paper follows from these aims. Section 2 characterizes frontier AI as a thought-experiment instrument and catalogues the main types of conceptual work it can support. Section 3 reviews how scientists across fields and career stages are currently using AI, emphasizing the gap between overt and covert practice. Section 4 examines gatekeeping and cognitive monopoly dynamics that shape attitudes toward AI-assisted thinking. Section 5 distinguishes legitimate concerns from protective rationalizations. Section 6 proposes

constructive norms for centaur practice, while Section 7 offers concrete templates that younger scientists might adapt in their own work. Section 8 considers institutional implications for training, evaluation, and the future division of cognitive labor. Section 9 concludes with reflections on the costs of prolonged taboo and the opportunities opened by normalizing AI-assisted thought experiments. Section 10 provides references supporting the analysis.

2. Frontier AI as a Thought-Experiment Instrument

To see frontier AI clearly in this context, it helps to stop thinking of it as "just a chat interface" and start treating it as a configurable instrument for exploring hypothetical structure. The visible surface is a text box, but behind that surface is a system that can generate, transform, and interrogate patterns in language, argument, and scenario design at a scale and speed that no individual human can match. In this sense, it belongs in the same family as microscopes and particle accelerators, but pointed inward, at the space of possible ideas rather than the space of observable matter.

Unlike a traditional instrument, however, this one is reflexive. It can be instructed to critique its own outputs, to adopt different roles, to maintain or vary assumptions across iterations, and to keep track of what has already been considered. These capacities make it well suited to the particular demands of thought experiments: the construction of stylized worlds, the systematic variation of conditions within those worlds, and the disciplined tracing of consequences.

The question, then, is not whether AI can be used for thought experiments, but how to characterize and constrain that use so that it becomes a stable, trustworthy component of scientific practice. This section describes the transition from chat to conceptual laboratory, surveys the main types of AI-assisted thought experiments across disciplines, identifies the safe region of purely ideational work, and clarifies what remains irreducibly human at the acceptance boundary.

2.1 From chat interface to conceptual laboratory

The most obvious way to encounter a frontier model is as a conversational partner. A scientist types a question; the model replies; the interaction proceeds turn by turn. In this mode, it feels natural to think of the system as a glorified search engine or a very fluent assistant. The conceptual potential of the interaction is easy to overlook.

The same underlying system can, however, be orchestrated as a small laboratory for structured reasoning. The scientist can ask the model to instantiate multiple roles simultaneously: one instance as an advocate of a hypothesis, another as a skeptic, a third as a methodologist focused on experimental design, a fourth as a historian of the field. The outputs of these roles

can then be fed into a fifth instance tasked solely with synthesizing points of agreement and unresolved tension.

Beyond role-play, the laboratory metaphor highlights the importance of control and logging. Prompts can be treated like experimental settings: specifying assumptions, constraints, and objectives, then observing how different configurations change the pattern of generated arguments or scenarios. Iterations can be saved and compared, just as one would compare successive runs of a simulation. Over time, a lab notebook of prompts, outputs, and human decisions can accumulate, providing a traceable record of how a thought-experiment space was explored.

Seen this way, the chat interface is only the front panel of a more general instrument. What matters is not the illusion of conversation, but the ability to configure and reconfigure the model's generative behavior to probe the structure of a problem from many angles, under explicit, documented conditions.

2.2 Types of AI-assisted thought experiments across disciplines

Because the medium is language, the same instrument can support very different kinds of thought experiments in different fields.

In the physical sciences, AI can be used to articulate counterfactual laws or parameter regimes and to reason through their qualitative consequences. A physicist might ask the model to imagine a world in which a fundamental constant has a different value, or in which a particular interaction term vanishes, and then explore what would have to change in the surrounding theoretical structure for this to be coherent. The model can generate candidate answers, list likely contradictions, and suggest observables that would distinguish such a world from ours.

In the life sciences, AI can help in designing stylized versions of biological systems or clinical trials. A biologist might propose several hypothetical regulatory architectures for a gene network and ask the model to describe their expected qualitative behavior under perturbation. A clinician might describe a trial design and ask the model to enumerate possible confounders, ethical dilemmas, or ways the trial could fail to detect a real effect. These are thought experiments about design, not instructions for actual patients.

In the social sciences, AI is particularly suited to scenario building. An economist might outline three simplified policy regimes and ask the model to project plausible chains of institutional response, public perception, and media framing over time. A political scientist might describe a stylized authoritarian system and ask the model to explore how different communication technologies would alter its stability or modes of collapse. The model's value here is not to predict the future, but to offer a wide range of structured possibilities against which human judgment can react.

In philosophy and ethics, AI can stage refined variants of familiar dilemmas and generate new ones. A moral philosopher can ask the model to take a particular principle and deliberately construct edge cases where that principle yields uncomfortable or conflicting prescriptions. A philosopher of mind can use the model to design self-consistency tests or behavioral criteria for ascribing understanding to an artificial system, then see how different theories survive those tests.

Even in mathematics and theoretical computer science, where formal proof remains the gold standard, AI can assist in the pre-formal stage. It can generate candidate counterexamples, unusual special cases, or analogies to better-understood structures in other areas. It can sketch heuristic arguments that a human then formalizes or rejects. In all these disciplines, the core pattern is the same: AI extends the human capacity to populate and inspect the space of hypotheticals before any formal or empirical commitment is made.

2.3 The "safe region": ideas, not actions or deployments

Not all AI use is alike in its risk profile. The regime considered in this paper is deliberately constrained to what might be called the safe region: the generation, transformation, and critique of ideas that remain within the realm of thought experiments and do not themselves constitute actionable plans for high-risk domains.

Within this region, the model is used to imagine worlds, articulate arguments, and expose assumptions. It does not output operational protocols for dangerous biological agents, detailed exploit code for critical infrastructure, or manipulative scripts targeted at specific vulnerable populations. Those categories belong to a different discussion, where alignment, dual-use concerns, and access control dominate.

The safe region is defined by two boundaries. Upstream, the inputs consist of high-level problem descriptions, existing theories, and stylized scenarios. They are not requests for concrete, step-by-step instructions in sensitive domains. Downstream, the outputs are filtered by the scientist's own judgment before they are allowed to shape experiments, policies, or public communication. The ideas that survive this filtration must still conform to ethical, legal, and disciplinary standards.

This does not mean that thought experiments are harmless in an absolute sense. Imagined scenarios can influence real-world attitudes and decisions, and bad reasoning at the conceptual level can misdirect whole research programs. But compared to domains where AI outputs directly interface with the physical world, the risks are slower, more diffused, and more familiar. They resemble the longstanding hazards of sloppy theorizing or persuasive but unsound argument, amplified by a powerful new generator.

Recognizing and respecting this safe region is essential. It allows scientists to be ambitious about using AI to expand their imaginative reach, while maintaining clear red lines around domains where additional safeguards or outright prohibitions are appropriate.

2.4 Human responsibility at the acceptance boundary

Even within the safe region, there remains a sharp boundary at which responsibility shifts decisively from machine to human. On one side lies the generation of candidate ideas, arguments, and scenarios by the model. On the other lies the act of acceptance: the moment a scientist decides to treat an AI-suggested construct as worthy of serious consideration, further development, or public endorsement.

The acceptance boundary is where scientific responsibility is concentrated. Models can propose endlessly, but they cannot decide which proposals to take seriously, how much weight to assign them, or what risks attend their adoption. Those judgments depend on background knowledge, ethical sensibilities, tacit field norms, and a sense of proportion that remain human capacities, even if they can be supported by tools.

A disciplined practice of AI-assisted thought experiments therefore centers on habits at this boundary. These include systematically cross-checking model-generated claims against primary sources, subjecting especially attractive ideas to extra scrutiny, and maintaining a record of which AI contributions have been adopted, modified, or rejected, and why. It also includes being explicit, at least in private notes, about when an argument or scenario originated with the model and when it emerged independently.

By sharpening the distinction between proposal and acceptance, scientists can harness the generative strengths of frontier AI without abdicating their own role. The model becomes a prolific source of raw material in the laboratory of ideas, but the scientist remains the one who decides what enters the architecture of belief and practice. That division of labor, properly understood and maintained, is what makes AI a thought-experiment instrument rather than a hidden author or an unacknowledged decision-maker.

3. How Scientists Are Actually Using AI Today

The public debate about AI in science often sounds as if the question were whether researchers will ever use these systems at all. On the ground, that question has already been answered. Across many disciplines, scientists are integrating AI into daily work in ways that are both mundane and profound. The real distinction is not between users and non-users, but between officially acknowledged uses that fit existing norms and unofficial uses that are kept quiet.

This section sketches that landscape. It begins with the visible surface: help with language, formatting, and boilerplate code. It then turns to the hidden layer of ideation and conceptual scaffolding that many scientists now conduct in private chat logs and notebooks. It notes differences across fields, career stages, and regions. Finally, it argues that these patterns amount to a double life: AI-rich practice paired with AI-poor narratives, and that this split is itself shaping the trajectory of scientific norms.

3.1 Surface adoption: editing, summarization, and boilerplate code

The most visible and non-controversial uses of AI in scientific work sit close to the editorial layer. Researchers feed drafts, abstracts, or emails into a model and ask for improvements in clarity, grammar, or structure. The model returns more polished prose, alternative phrasings, and suggestions for smoothing transitions. This use case is widely accepted, especially for non-native speakers who are acutely aware of the penalties attached to awkward language in peer review and grant competitions.

Summarization is a close second. Scientists paste in an abstract or manuscript and request a concise overview, a lay summary, or a bullet-point list of main findings. In some cases they ask for structured outputs, such as summaries focused specifically on methods, limitations, or potential applications. For overloaded readers, this accelerates triage: deciding which papers to read in full and which to skim or set aside. For teaching and supervision, it supports the creation of reading guides and discussion prompts.

A related, now almost routine use is code assistance. Models are asked to generate boilerplate scripts in common languages, translate between programming environments, or refactor existing code for readability. In computational fields, this has become analogous to using a search engine plus online forums, but with a more interactive loop. Scientists lean on AI to handle routine tasks, freeing their time for more specialized work.

These uses are easy to justify under existing norms. They resemble support from a copy editor, a junior collaborator, or a programming-savvy colleague. They do not overtly challenge the idea that the senior scientist remains the primary source of conceptual innovation. As a result, they are the ones most often openly acknowledged in papers and lab practices.

3.2 Hidden adoption: private notebooks and unacknowledged ideation

Beyond the surface, many scientists now engage with AI as a conceptual partner in ways that are rarely disclosed. In private chat histories and personal notebooks, they ask models to help brainstorm hypotheses, generate alternative explanations for puzzling results, or design variants of experimental setups. They may request lists of potential confounders, critiques of their own arguments, or thought experiments that push their intuitions to the limit.

Sometimes this takes the form of structured brainstorming: “Give me ten distinct mechanisms that could explain this observation, assuming nothing beyond standard theory.” The scientist then sifts through the list, discards most, and develops a few further. Sometimes it is interactive refinement: “Here is my hypothesis; attack it as a hostile reviewer would,” followed by back-and-forth where the model raises objections and the scientist responds or adjusts.

Another hidden use is outline generation. Researchers describe a project or idea in rough terms and ask the model to sketch possible structures for a paper, grant proposal, or talk. The AI suggests sections, ordering, and transitions; the human then revises or repurposes these outlines. Over time, this can significantly shape how arguments are framed and what appears as the natural structure of a piece of work, even if no AI involvement is acknowledged.

These ideational uses shade into ghostwriting. In some cases, paragraphs or even whole sections of text are drafted with substantial model input, then edited by the scientist. The final result may be indistinguishable in style from human-only writing, making detection difficult. Yet because norms and policies around conceptual AI assistance are unsettled, many researchers treat this as something best not mentioned, especially when the AI’s contribution extends beyond spelling and grammar.

The net effect is that a growing fraction of the conceptual landscape is being shaped in dialogue with AI systems, but official documentation presents these developments as if they emerged solely from human minds. The private laboratory of ideas includes AI; the public narrative does not.

3.3 Differences across fields, career stages, and geographies

Patterns of AI use are not uniform. They vary across disciplines, career stages, and regions in ways that reflect both technical needs and institutional cultures.

Fields that already sit near computation and text, such as computer science, bioinformatics, and parts of quantitative social science, have generally adopted AI tools more quickly. In these areas, using a model to generate code snippets, debug, or sketch algorithms fits naturally within existing workflows. Conceptual uses, such as asking AI to suggest variants of model architectures or optimization strategies, also appear earlier because the boundaries between code and ideas are already porous.

In contrast, in experimental wet-lab biology, clinical medicine, or highly formal branches of mathematics, uptake is more uneven. Some researchers embrace AI for literature triage and writing assistance but remain cautious about relying on it for conceptual moves, either because they distrust its technical competence or because their communities are more conservative about what counts as legitimate reasoning. In some of these fields, the cost of being seen as

over-dependent on AI can be high, especially for early-career scientists seeking tenure or clinical credibility.

Career stage matters as well. Younger researchers, particularly graduate students and postdocs, tend to be more willing to experiment with AI in all parts of their workflow. They have grown up with digital tools and face acute pressure to produce more with fewer resources. For them, AI can feel like a necessary prosthesis just to keep up. At the same time, they are more vulnerable to gatekeeping. Senior scientists control letters of recommendation, hiring, and promotion, and may negatively judge visible reliance on AI for conceptual work.

Senior scientists are, on average, more cautious and selective. Some adopt AI intensively in private, using it to maintain breadth across a sprawling literature, explore unfamiliar methods, or compensate for limited time. Others avoid it almost entirely, either out of distrust or because they feel they do not need it. Where senior figures are openly supportive, labs often develop explicit norms for AI use. Where they are skeptical, junior members may resort to using AI quietly, without institutional recognition.

Geography and resource distribution also play a role. In settings with high access to commercial frontier models, adoption can be rapid but policy sensitive, especially where funders or ethics boards issue guidelines. In regions with limited access due to cost, connectivity, or regulatory restrictions, researchers may rely on smaller open models or local deployments, with different strengths and limitations. For scientists in resource-constrained environments, AI can offer a disproportionate benefit by providing access to coding help, language polishing, and conceptual dialogue that would otherwise be unavailable. At the same time, unequal access risks widening existing disparities in who can effectively participate in fast-moving fields.

These variations complicate any simple story about “how scientists use AI.” Instead, they reveal a patchwork of practices shaped by technical demands, institutional norms, and power relations within and between disciplines and regions.

3.4 The emerging double life: AI-rich practice, AI-poor narratives

Taken together, surface and hidden uses produce an emerging double life in scientific practice. On one side is the lived reality: many researchers routinely rely on AI systems to brainstorm, critique, and refine ideas, to structure arguments, and to handle an increasing share of textual and computational labor. On the other side is the official story as told in methods sections, author guidelines, and professional norms, where AI is at most a peripheral tool for polishing language and code.

This mismatch has several consequences.

First, it distorts the record of how science is actually done. Future historians and methodologists reading the literature will see a sanitized account that omits an important component of the cognitive ecology in which ideas emerged. The trail from question to result will appear more linear and more human-centric than it really was, hiding the iterative, AI-mediated exploration of possibilities that increasingly lies upstream.

Second, it creates asymmetric risks for different groups. Senior scientists who use AI privately while discouraging public acknowledgment can benefit from its accelerations without jeopardizing their reputations, because their established status insulates them. Junior scientists, by contrast, face a dilemma. If they openly disclose extensive AI assistance, they may be seen as less original or independent. If they conceal it, they participate in a culture of quiet misrepresentation and may be punished later if norms shift or disclosure becomes mandatory.

Third, the double life slows the development of robust, field-specific norms. As long as conceptual AI use remains an open secret rather than an openly discussed practice, communities struggle to articulate clear expectations about disclosure, responsibility, and credit. Debates happen in the abstract, often framed as if the main issue were future hypothetical uses rather than current realities. Policies are drafted without full knowledge of how deeply AI has already penetrated the day-to-day life of research.

Finally, the split reinforces the very hierarchies that feel threatened by AI. By restricting official legitimacy to low-level uses and stigmatizing or ignoring higher-level ones, gatekeepers preserve existing notions of who counts as a “real thinker.” They can quietly draw on AI themselves while publicly insisting that true originality resides only in unassisted human minds. This narrative helps maintain cognitive monopoly and status structures, at the cost of honesty about how contemporary science actually proceeds.

Recognizing this double life is a necessary step toward resolution. The choice is not between a pristine, AI-free ideal and a fully automated future. It is between a scientific culture that acknowledges and governs AI-assisted thought experiments openly, and one that continues to rely on them in practice while denying their role in public. The next sections argue that the former path, though disruptive, offers a more stable and ethically coherent way to integrate frontier AI into the life of inquiry.

4. Gatekeeping and Cognitive Monopoly Protection

The resistance to AI-assisted thought experiments is often framed in abstract terms: worries about rigor, originality, or the integrity of the scientific record. Beneath those themes lies a

simpler and more personal story. For decades, senior scientists have occupied a cognitive monopoly. They were the ones who could read widely, synthesize quickly, write forcefully, and orchestrate complex debates in their heads. That bundle of skills justified their authority over careers, funding, and the direction of fields.

Frontier AI erodes this monopoly. It gives younger and less-resourced researchers access to some of the same cognitive leverage that once belonged only to a small elite. The result is a pattern of gatekeeping that aims, consciously or not, to preserve existing hierarchies. This section examines how senior cognitive advantages are being challenged, how narratives about "real thinking" are used to police AI use, how labor and credential structures shape AI discourse, and how these forces translate into pressure on younger scientists.

4.1 Senior cognitive advantages and how AI erodes them

In the pre-AI environment, senior scientists typically held three interlocking advantages.

First, they had breadth. Years of reading across subfields and neighboring disciplines allowed them to see connections that younger colleagues could not. They could draw analogies, recall obscure precedents, and place new findings into a wide conceptual landscape.

Second, they had depth of argument. Experience in writing grants, papers, and reviews sharpened their ability to construct and dismantle complex lines of reasoning. They knew where reviewers would attack, which assumptions were fragile, and how to pre-empt objections.

Third, they had fluency in expression. They could turn ideas into polished prose quickly, giving their work an air of inevitability and professionalism that carried weight in committees and panels.

Frontier AI challenges all three advantages at once. A junior researcher with a model at hand can rapidly map out literatures they have not yet read in full, generate analogies across fields, and obtain preliminary critiques of their own arguments. They can ask the model to play the hostile reviewer, to stress-test a conceptual framework, or to suggest alternative structures for a paper. They can also polish their language and presentation to a level that narrows the visible gap between their writing and that of established figures.

This does not mean that AI renders experience irrelevant. Senior scientists still have tacit knowledge, judgment, and field-specific intuition that models cannot replicate. But it does mean that the margin by which experience dominates early-career talent becomes smaller. The distance between "raw" and "refined" thinking shrinks when refinement can be outsourced in part to a machine.

For those whose identity and authority are heavily invested in that distance, the change is unsettling. What once felt like a natural consequence of time and effort now looks, in part, like a

positional privilege that can be technologically undercut. The temptation is strong to delegitimize the tools that narrow the gap, or at least to confine them to domains that do not touch core cognitive work.

4.2 Narrative control over what counts as "real thinking"

Gatekeeping rarely announces itself as self-protection. Instead, it presents as a defense of standards. One of the most powerful tools in this defense is narrative control over what counts as "real thinking" in science.

For much of modern scientific history, the canonical story has been one of individual or small-group genius. Ideas are imagined as emerging from the minds of gifted individuals, perhaps in conversation with colleagues and the literature, but fundamentally authored by humans. Instruments and tools are important, but they are accessories. The real work of discovery happens in the head.

This story is deeply ingrained in training, in popular accounts of science, and in the self-understanding of many senior researchers. It also justifies their role as arbiters. If ideas live primarily in minds, and some minds have demonstrated their worth over decades, then it is natural to grant those minds authority over what counts as a good question, a valid argument, or a promising direction.

AI-assisted thought experiments destabilize this narrative. When a large language model contributes materially to the generation, refinement, and organization of ideas, the clean line between "tool" and "thinker" blurs. The model is not conscious or responsible, but it is also not just a passive instrument. Its generative patterns shape which possibilities are easy to see and which remain hidden. It becomes part of the cognitive environment in which human thinking happens.

To preserve the older narrative, gatekeepers often draw a sharp distinction between acceptable and unacceptable AI use. Help with grammar, formatting, and boilerplate code is permitted because it fits the image of tools as peripheral. Use for ideation, argument structure, or hypothesis generation is treated as suspect because it challenges the notion that ideas originate wholly within human minds.

This distinction is then encoded in policy and informal commentary. Editorials caution against "relying" on AI for creativity. Guidelines insist that authors remain responsible for conceptual content while providing little guidance on how to acknowledge AI's role. Senior voices warn that using AI for "real thinking" undermines independence. In each case, the message is the same: true scientific thought is what happens without the machine, even as the machine quietly saturates everyday practice.

By defining “real thinking” in ways that exclude AI-assisted processes, gatekeepers protect the symbolic capital of unassisted cognition. They can continue to claim a special status as bearers of human originality while allowing AI to infiltrate less prestigious layers of work.

4.3 Labor, credential, and status preservation in AI discourse

Beyond cognitive monopoly and narrative control lies a more straightforward interest: the preservation of labor structures and credentials that support existing careers.

Academic science is built on pyramids. A small number of senior, secure positions rest atop a larger mass of precarious labor: graduate students, postdocs, adjuncts, and contract staff. Advancement depends on producing high volumes of work, securing grants, and satisfying evaluators who themselves came up through a pre-AI system. The value of advanced credentials is tied to the idea that the skills they certify are scarce and hard to acquire.

AI tools threaten to disrupt this equilibrium in several ways.

If a single researcher, equipped with frontier AI, can do the conceptual and writing work that previously required a small team, the implicit need for as much junior labor may decline. If AI can help a motivated outsider produce work that rivals that of someone with formal training, the monopoly value of traditional credentials may weaken. If AI can compress the time needed to move from novice to contributor, career timelines may need to be rethought.

Faced with these possibilities, institutions and individuals for whom the current structure works reasonably well are understandably hesitant. Few are eager to acknowledge that the training they provide could, in some dimensions, be accelerated or partially bypassed. Few are keen to admit that their own productivity is now heavily dependent on tools available to many others.

As a result, AI is often framed in discourse as a potential threat to jobs, especially those of junior scientists, but rarely as a force that might just as well reduce the necessity of certain senior roles or administrative burdens. The conversation about job protection focuses downward, on laboratories and staff, rather than upward, on the distribution of authority and reward.

In parallel, credentialing bodies and professional societies move cautiously. They recognize the need to update curricula to include AI literacy, but they also resist shifts that would dilute the centrality of traditional skills. Assessments and examinations often prohibit or restrict AI use even in contexts where, outside the exam room, AI is ubiquitous. The implicit message is that “real competence” is what remains when AI is removed, even if the future workplace will rarely remove it.

This labor and credential preservation shows up in how AI is discussed in policy documents, hiring criteria, and promotion dossiers. Contributions that leverage AI in visible ways may be devalued relative to those that present as purely human, even if the latter quietly rely on similar

tools. Scientists learn quickly what kinds of work are rewarded and adjust their behavior accordingly, reinforcing a system in which AI is both relied upon and rhetorically marginalized.

4.4 Peer and peer-review pressures on younger scientists

The combined effect of cognitive monopoly instincts, narrative control, and labor preservation falls most heavily on younger scientists. They inhabit a professional world where AI is practically indispensable and socially ambiguous. Their peers, advisors, and reviewers send mixed messages about what is allowed and what is wise to disclose.

At the level of everyday lab life, informal comments from supervisors and colleagues set the tone. Some mentors encourage careful, explicit use of AI, treating it as a new instrument to be mastered. Others issue blanket warnings: do not let AI touch your ideas, do not trust AI summaries, do not mention AI in your methods. These warnings may be framed as concern for rigor or fear of plagiarism accusations, but they also reflect discomfort with changing status hierarchies.

Peer review amplifies these pressures. Reviewers who are skeptical of AI may look for signs of its use and interpret them negatively, questioning the originality or independence of the work. Even when no explicit mention is made, a manuscript that reads in a certain way or uses certain phrase patterns may trigger suspicion. Authors, aware of this possibility, err on the side of under-reporting AI involvement to avoid jeopardizing acceptance.

Grant review panels can behave similarly. A proposal that foregrounds AI-assisted conceptual work may be seen as risky or unserious compared to one that presents human cognition as the sole engine of innovation. Junior applicants quickly learn that it is safer to describe AI only as a tool for data management or analysis, not as a partner in hypothesis generation or theoretical structuring.

Within peer groups, the dynamics are more mixed. Younger scientists share tips and prompts, swap successful AI workflows, and commiserate over the awkwardness of hiding how much they rely on these systems. At the same time, they may judge one another for perceived over-dependence, internalizing senior narratives about “real thinking.” The result is a culture in which many people are using AI aggressively in private while maintaining a posture of distance in public.

All of this creates a double bind. To be competitive, younger scientists almost have to use AI for more than surface tasks; the volume and complexity of demands make it difficult to keep up otherwise. But to be seen as properly independent and original, they must act as if AI were marginal to their thinking. The safest path is to employ AI heavily in invisible ways and to avoid experimental practices that would require open acknowledgment of AI’s conceptual contributions.

In such an environment, the promise of AI-assisted thought experiments is realized in a partial and distorted fashion. The tools are there; the imaginative uses are known; but the social field penalizes the very transparency that would allow robust norms to emerge. The next sections argue that it is possible to break this pattern without surrendering rigor or flooding the literature with unexamined machine-generated ideas, by articulating positive standards for centaur practice and concrete templates for disclosure.

5. Legitimate Concerns Versus Protective Rationalizations

Not every objection to AI-assisted thought experiments is self-serving. Some worries are serious and grounded in real properties of current systems. Models do, in fact, hallucinate, overfit to patterns in their training data, and present speculative structure with a confidence that can mislead even experienced readers. It is also possible to lean on AI so much that one's own skills atrophy, or to blur authorship and accountability in ways that corrode trust.

At the same time, those genuine concerns can be stretched and repurposed. They can be used to justify blanket prohibitions in settings where more nuanced safeguards would suffice, or to stigmatize uses of AI that are no more risky than other, long-accepted practices. Distinguishing between real dangers and hierarchy-preserving myths is essential if scientific communities are to govern AI wisely rather than reactively.

This section unpacks four clusters of concern: epistemic slippage and hallucinated structure, fears of shallow dependence and training loss, accountability and authorship in AI-shaped idea pipelines, and the line between legitimate caution and protective rationalization.

5.1 Epistemic slippage and the problem of hallucinated structure

One of the most widely cited worries about large language models is that they generate confident, articulate, and sometimes entirely fabricated content. This may involve nonexistent references, misquoted results, or the subtle reweaving of disparate facts into a narrative that feels coherent but has no firm empirical foundation. In the context of thought experiments, the risk is less about discrete false statements and more about what could be called hallucinated structure.

Hallucinated structure arises when a model offers a neatly organized conceptual framework, a hierarchy of mechanisms, or a space of scenarios that appears internally consistent but is only loosely anchored to the underlying literature or data. Because the presentation is polished and the categories feel familiar, it can be easy for a human reader to accept this structure as a map of the real conceptual terrain rather than as a tentative proposal.

In the laboratory of thought experiments, this manifests in several ways. A model might propose a taxonomy of possible explanations for a phenomenon that subtly excludes important alternatives. It might generate a seemingly exhaustive list of confounders that misses the most dangerous one. It might reframe a moral dilemma in terms that bias the reader toward a particular resolution. In each case, the danger is that the human scientist will mistake generative fluency for epistemic authority.

This concern is legitimate. It points to the need for disciplined practices that keep the distinction between proposal and acceptance sharp. Mitigations include systematic cross-checking of model-suggested claims against primary sources, explicit notation of which structures emerged from AI interactions, and deliberate efforts to break and stress-test model-generated frameworks before adopting them. The presence of hallucinated structure does not rule out AI-assisted thought experiments, but it does demand a culture of skepticism proportionate to the model's persuasive power.

5.2 Fears of shallow dependence and lost training value

Another recurring worry is that ready access to AI will produce a generation of scientists who can assemble the appearance of sophisticated reasoning without having internalized its substance. If models can draft arguments, generate counterexamples, and suggest experiment designs, will trainees still learn how to do these things themselves? Will they become dependent on prompts instead of developing their own capacity for deep, autonomous thought?

This concern echoes earlier debates about calculators in mathematics, reference managers in writing, and software for numerical simulation. In each case, there were real risks of shortcuts eroding foundational skills. In each case, the long-term outcome depended heavily on pedagogy. Where educators simply allowed tools to replace basic practice, competence suffered. Where they strategically integrated tools after certain thresholds of understanding were reached, overall capacity increased.

With AI, the stakes are higher because the tools operate closer to the core of conceptual work. There is a plausible path in which students learn to query models for explanations, analogies, and arguments without ever grappling with the friction of confusion that traditionally forces deeper integration. The danger is not just that they will be less skilled in certain techniques, but that they will lack the mental habits of sustained, independent wrestling with a problem.

This, too, is a legitimate concern. It argues for carefully designed no-AI zones in training: stages or assignments where students are explicitly required to think and write without model assistance, and where the goal is to build their own internal repertoire. It also suggests that AI literacy should include meta-skills: knowing when to withhold the tool, how to use it for

exploration rather than outsourcing, and how to recognize the difference between borrowed structure and owned understanding.

The existence of these risks does not justify banning AI from conceptual spaces altogether, any more than the risk of dependence justified banning calculators from all mathematical practice. It does justify curriculum design that treats AI as a powerful but potentially distorting amplifier, to be introduced in ways that strengthen rather than weaken the underlying cognitive muscles.

5.3 Accountability and authorship when AI enters idea pipelines

A third cluster of concern centers on responsibility. When AI systems are involved in generating hypotheses, structuring arguments, or designing thought experiments, who is accountable for errors, omissions, or harms that trace back to those contributions? How should credit and authorship be assigned when some of the conceptual scaffolding was proposed by a model?

On the accountability side, the answer is conceptually clear but practically demanding. Models cannot bear responsibility. They do not have intentions, do not understand the stakes of their outputs, and cannot be sanctioned or persuaded. Responsibility for any idea that enters the public record or guides action lies with the human authors who choose to endorse and propagate it. If an AI-assisted thought experiment leads to a misguided research program or a harmful policy, the fault rests with those who failed to interrogate and constrain the machine's proposals.

Operationalizing this principle requires transparency about how AI was used and robust internal review processes. When model-generated constructs play a significant role in shaping a paper or proposal, authors should be able to say, at least to themselves and their collaborators, which parts arose from human intuition and which from machine suggestions. They should also be prepared to defend the steps by which they evaluated and accepted those suggestions.

Authorship raises different questions. If AI contributes materially to the generation and organization of ideas, some may feel tempted to treat it as a co-author. This impulse is understandable but misplaced. Authorship is not just a record of contribution; it is a marker of responsibility, accountability, and the capacity to respond to critique. Assigning authorship to entities that lack these capacities dilutes the meaning of authorship and confuses lines of responsibility.

A more appropriate approach is to treat AI as an instrument whose role is described in methods, acknowledgments, or separate documentation, while reserving authorship for humans. Within that human set, however, AI can complicate the distribution of credit. Junior researchers who use models to amplify their conceptual reach may be perceived as less original, even when the critical moves in evaluating and integrating AI outputs are theirs. Senior

researchers who quietly rely on AI may receive disproportionate credit because their historical reputation masks the tool's contribution.

These tensions will not be resolved by simple prohibitions or declarations. They require communities to revisit authorship criteria, to value the work of orchestrating and scrutinizing AI contributions, and to ensure that accountability remains clearly human even as the texture of idea generation changes.

5.4 Distinguishing real risks from hierarchy-preserving myths

Legitimate concerns about hallucination, dependence, and accountability coexist with less grounded narratives that primarily serve to protect existing hierarchies. Myths of this kind often use the language of rigor while targeting uses of AI that are no more risky than accepted practices, or they deploy worst-case scenarios from one domain to justify blanket bans in another.

One common pattern is to point to notorious examples of AI failure in high-stakes settings and then treat any use of AI in thought experiments as morally adjacent. This ignores the crucial difference between systems whose outputs directly control actions in the world and systems whose outputs feed into human scrutiny long before they have any external consequences. Another pattern is to insist that "real thinking" cannot involve AI at all, while quietly accepting the use of equally opaque human intuition or informal discussions that never make it into the record.

A useful way to separate risk from rationalization is to ask a set of diagnostic questions. Does the concern persist if AI use is fully disclosed and subject to cross-checking? Is the proposed restriction targeted at specific failure modes, or does it amount to a generalized stigma against conceptual assistance? Are comparable risks tolerated when they arise from human collaboration, rule-of-thumb reasoning, or disciplinary tradition? Does the stance fall symmetrically on senior and junior practitioners, or does it disproportionately constrain those with less power?

When concerns are specific, symmetrical, and tied to clear mechanisms of harm or distortion, they are likely to be legitimate. When they are vague, asymmetrical, and selectively applied, they are more likely to function as protective myths. Scientific communities do need defenses against degraded reasoning and irresponsible use of powerful tools. They do not need defenses against the erosion of unjustified cognitive monopolies or the discomfort of shared authorship in the broad sense of collective thought.

The task, then, is not to dismiss all worries about AI as conservatism, nor to accept all restrictive impulses as prudence. It is to cultivate a culture that can name both the real dangers and the

self-serving stories, and that can design norms which protect the integrity of inquiry without freezing the distribution of cognitive power in its current, historically contingent form.

6. Centaur Practice: Positive Norms for AI-Assisted Thought Experiments

If the problem is not that AI-assisted thought experiments are impossible or inherently unsafe, but that their use is socially fraught, then the natural next step is to articulate what good practice would look like. Rather than framing AI as something that must either be excluded from thinking or allowed to overrun it, we can treat it as a component in a composite intelligence. The question becomes how to configure that composite so that it enhances, rather than erodes, the qualities that make scientific reasoning trustworthy.

The centaur metaphor is useful here. A centaur is neither a human wearing a machine nor a machine disguised as a human, but a hybrid in which each part has a distinct role. In scientific centaur practice, the machine contributes breadth, speed, and diversity of proposals; the human contributes judgment, responsibility, and the ability to integrate ideas into a wider ethical and empirical context. This section defines that configuration more precisely, proposes principles for legitimate AI use in conceptual work, suggests ways to be transparent without turning AI into an object of fascination, and considers how to evaluate originality when machine scaffolding is present.

6.1 Defining the centaur: human judgment plus machine breadth

In the context of thought experiments, a centaur is a stable role configuration, not a vague partnership. It can be characterized by three elements.

First, the division of labor. The machine's primary role is generative: producing candidate hypotheses, alternative framings, critical questions, and structured scenario variations. It works best when asked to enumerate possibilities, test consistency, or simulate how different perspectives might respond. The human's role is selective and integrative: deciding which of these outputs are worth pursuing, how they fit with existing knowledge, and what ethical and practical implications they have.

Second, the direction of control. The human sets goals, constraints, and stopping conditions. They decide when to invoke the machine, what information to feed it, and how to interpret its responses. The model does not autonomously decide which problems to tackle or which conclusions to endorse. It is more like a configurable lens than an independent agent.

Third, the locus of responsibility. In centaur practice, responsibility for every believed claim, published argument, and proposed experiment remains human. The model may influence the trajectory of thought, but it cannot share accountability. If a centaur configuration yields a bad

idea, the correct response is not to blame the machine but to interrogate the human choices about how the tool was used and where acceptance boundaries were set.

Defining the centaur in this way clarifies that the goal is not to simulate “pure” human thinking or to celebrate machine autonomy. It is to design a composite system in which each part contributes its strengths while known weaknesses are mitigated by the other. That design work can itself become explicit and teachable.

6.2 Principles for legitimate AI use in conceptual work

From this definition, a set of practical principles follows. They do not resolve every edge case, but they sketch a baseline for legitimate AI use in thought experiments.

One principle is proposal, not proof. AI contributions should be treated as proposed structures, not as evidence or authority. A model may suggest that a certain mechanism, analogy, or scenario is worth considering. It cannot establish that mechanism as real, that analogy as correct, or that scenario as likely. Human validation is always required, whether through literature review, formal derivation, or empirical test.

A second principle is contestability. Any idea that enters the scientific record should be open to challenge, including challenges that explicitly target its AI-mediated origins. If a model-generated taxonomy or argument becomes central to a paper, authors should be prepared to show how it withstands scrutiny by people who did not participate in the original centaur interaction.

A third principle is proportional skepticism. Because models are known to hallucinate and to over-extend patterns, their outputs should be subjected to at least as much scrutiny as analogous suggestions from an enthusiastic but fallible human collaborator. The more consequential a conceptual move is for a project, the more carefully its AI-mediated lineage should be examined.

A fourth principle is scope limitation. Certain domains, such as detailed protocols for dual-use biological work or exploit strategies for critical infrastructure, may be placed off-limits for AI assistance in all forms, including thought experiments. This is not because thought experiments themselves are dangerous, but because the line between hypothetical and operational content can blur in these areas. Explicit red lines, agreed upon at the lab or disciplinary level, help prevent drift from safe conceptual space into inadvertent instruction.

A fifth principle is reciprocity. If a community benefits from AI in conceptual work, it should also take on the responsibility of contributing to better understanding and governance of these tools. That might include sharing anonymized prompts and workflows for method development,

participating in evaluations of model behavior in their domain, or engaging with ethicists and policymakers trying to set reasonable rules.

Taken together, these principles reflect a stance that is neither naive adoption nor reflexive rejection. They treat AI as a powerful, noisy generator whose outputs can enrich thought if, and only if, they are disciplined by scientific norms.

6.3 Transparency without fetish: describing AI as method, not magic

If AI is to become a normal part of thought-experiment practice, it needs to be described in ways that are neither euphemistic nor theatrical. Overstatement, in which AI is presented as a quasi-mystical partner, invites backlash and muddies responsibility. Understatement, in which AI is buried under vague acknowledgments, prevents others from understanding and evaluating how it was used.

A constructive middle path is to describe AI in the same tone and location as other methods. In the methods section of a paper, authors can specify that they used a particular class of model to generate alternative hypotheses, enumerate possible confounders, or stress-test arguments, under stated prompts and constraints. They can note that these outputs were then screened, revised, or rejected by the human team. If relevant, they can mention how many iterations were run or how disagreements between model and human judgment were resolved.

This kind of transparency does not require full transcripts or technical specifications beyond what a reader needs to assess the role of AI in the work. It does not elevate the model to co-author or protagonist. It situates it as part of the apparatus through which the thought-experiment space was explored, alongside more familiar apparatus such as simulations, analytic calculations, or experiment design tools.

At the same time, transparency without fetish means not making AI the story when it is not. If a paper's main contribution is a new theoretical insight, an empirical finding, or a conceptual synthesis, and AI-assisted thought experiments played only a supporting role, there is no need to foreground the tool in the title or abstract. A brief, accurate methods description suffices. Overemphasizing AI can create perverse incentives, rewarding spectacle over substance.

For younger scientists, this style of description can also function as a shield. By treating AI as one tool among many, described in the same understated way as others, they signal both honesty and a refusal to make their own sophistication dependent on appearing either entirely unassisted or ostentatiously high-tech. It becomes possible to be frank about AI use without inviting accusations of either laziness or techno-theatricality.

6.4 Evaluating originality when some scaffolding is machine-generated

Perhaps the hardest cultural adjustment in centaur practice is redefining originality. If a model helps generate the scaffolding of an argument or the preliminary structure of a thought experiment, what does it mean to say that a scientist has had an original idea? How should committees and reviewers evaluate contributions that emerged from AI-assisted processes?

One starting point is to distinguish between raw proposal space and selective integration. Models can, in principle, enumerate vast numbers of possible hypotheses, analogies, or scenarios. The mere fact that a specific element can be generated by a model does not make it trivial, any more than the fact that many chemical compounds can be synthesized makes all molecules equally unoriginal. Originality in the centaur setting lies more in the choice, combination, and refinement of elements than in their brute proposability.

Consider a researcher who uses AI to generate twenty possible conceptual framings of a problem, then chooses one, modifies it significantly, connects it to a body of literature, and develops a set of precise thought experiments and testable implications. The model contributed scaffolding, but the creative work of selection and integration was theirs. Another researcher might ask a similar model for framings, accept the first suggestion uncritically, and build without serious interrogation. The visible outputs might look similar at a glance, but the underlying originality and responsibility differ.

Evaluation practices need to recognize this difference. Reviewers and committees cannot see inside every centaur interaction, but they can attend to how thoroughly ideas are developed, how well they connect to existing knowledge, and how robustly they withstand critique. Work that consists of loosely stitched model outputs will tend to be shallow and brittle. Work that uses AI-generated scaffolding as a starting point for deep engagement will show the same hallmarks of originality as pre-AI work: unexpected yet coherent connections, sensitivity to counterarguments, and a sense of necessity in the final structure.

Over time, it may also become useful to develop explicit expectations. For example, in training, students might be asked to submit both AI-assisted and unassisted work, demonstrating that they can operate in both modes. In hiring and promotion, candidates might be evaluated on their ability to explain their own thinking process, including how they orchestrate tools, rather than on a romantic ideal of solitary genius.

It is also worth noting that originality has never been a purely individual property. Scientific ideas emerge from conversations, seminars, conferences, and collective traditions. AI-assisted thought experiments simply add another layer to this already distributed process. Recognizing that does not diminish individual contributions; it situates them more honestly within a network of influences, of which models are now a part.

In the end, evaluating originality in a centaur world means asking a slightly different question. Not “did this idea arise in isolation within one mind?” but “how did this person work with the materials available, human and machine, to produce something that advances understanding?” The answer will still vary in quality, and not every use of AI will be worthy. But the measure shifts from purity of origin to depth and responsibility of use.

7. Practical Templates for Younger Scientists

The abstract arguments about centaur practice only help if they can be translated into concrete moves that a younger scientist can actually make without derailing a fragile career. The aim of this section is therefore pragmatic. It offers examples of how to describe AI-assisted thought experiments in a methods section, sketches disclosure patterns that are honest but not self-destructive under current norms, suggests lab-level charters for centaur use, and outlines strategies for responding to reviewer objections.

None of these templates is universal. Fields, journals, and institutions differ. But they illustrate a stance: normalize AI as a method, keep responsibility clearly human, and avoid both concealment and theatrical self-promotion. For a younger scientist navigating asymmetric power, even small shifts in how AI use is framed can make the difference between suspicion and acceptance.

7.1 A model “AI-assisted thought experiments” methods subsection

One barrier to honest disclosure is simply the lack of a familiar pattern. Authors know how to describe simulations, surveys, or bench protocols, but not how to describe a process that lives in prompts and conversation logs. A simple, neutral template can lower that threshold.

A generic methods subsection might look like this:

In addition to conventional literature review and internal group discussion, we used a large language model as a tool to support the design and refinement of thought experiments related to [topic]. Specifically, we conducted a series of structured prompt–response sessions in which the model was asked to:

1. Generate alternative hypotheses and explanatory mechanisms consistent with the empirical constraints described in Section X.
2. Enumerate potential confounding factors, boundary conditions, and failure modes for the proposed scenarios.
3. Adopt adversarial perspectives (for example, a skeptical reviewer) to identify weaknesses, missing assumptions, or underdeveloped lines of argument.

All model interactions were conducted by the authors. Outputs were treated as proposals rather than results. We screened, modified, or rejected model suggestions based on domain knowledge, consistency with the existing literature, and internal critique. No conclusions in this paper rest solely on model-generated content without independent human evaluation. The model was not used to generate or alter empirical data, perform statistical analyses, or produce any operational protocols.

Prompts and representative excerpts of model outputs relevant to the conceptual development of the paper are available from the authors upon reasonable request.

This template does several things at once. It locates AI alongside other methods, specifies what it was and was not used for, emphasizes human control at the acceptance boundary, and avoids dramatic language. It also makes room for variation. A more detailed version might distinguish between phases (initial brainstorming, refinement, stress-testing), or specify that multiple models or configurations were used. A more concise version might compress the description into a short paragraph if space is limited.

The point is not to impose a rigid format, but to demonstrate that AI-assisted thought experiments can be described in the same cool, precise tone as other parts of the method, without either hiding or mystifying the tool.

7.2 Disclosure patterns that are honest but not self-sabotaging

Under current conditions, younger scientists often worry that any mention of AI will lead reviewers or evaluators to downgrade their independence or originality. At the same time, they may feel uneasy about saying nothing when AI played a real role in shaping the work. Navigating that tension requires disclosure patterns that are honest, concrete, and calibrated to what readers actually need to know.

One pattern is function-focused disclosure. Instead of emphasizing the model's presence, authors emphasize its role. For example:

We used a large language model to assist in generating initial lists of possible mechanisms, which were then screened and developed by the authors.

We used a large language model to propose alternative framings of the central question; the final structure of the argument reflects the authors' selection and synthesis.

This kind of statement avoids both extremes. It does not pretend that ideas emerged from nowhere, but it also does not attribute agency or creativity to the model. It keeps the focus on what the human authors did with the suggestions.

Another pattern is tiered disclosure. The main text contains a succinct methods mention, while an appendix or supplementary file provides more detail for interested readers. For example, the body might say:

Conceptual development was supported by structured interactions with a large language model, described in detail in Supplementary Methods.

The supplement can then list prompt templates, interaction protocols, and examples of how AI-generated material was evaluated or modified. This separation keeps the main narrative uncluttered while preserving transparency for those who care about the AI component.

A third pattern is cumulative disclosure at the level of a research program rather than an individual paper. A scientist might include in their personal statement, lab website, or grant biosketch a clear description of how AI fits into their general workflow, even when individual papers only mention it briefly. That might look like:

In my group, we routinely use large language models as tools for hypothesis generation, critical questioning, and the design of thought experiments. All AI-generated suggestions are subjected to internal review and are not adopted without independent justification.

This kind of statement signals a stable practice and can serve as a reference point if questions arise about individual pieces of work.

The common thread is that disclosure is framed as a mark of conscientiousness, not confession. The message becomes: we use modern tools, we use them carefully, and we are prepared to explain how.

7.3 Lab-level norms: group charters for centaur use

Individual early-career scientists are often at the mercy of their immediate environment. A supportive supervisor can normalize centaur practice; a skeptical one can make it risky even to ask. One way to improve this local ecology is to articulate lab-level norms in a short charter or policy on AI use.

Such a charter might include elements like the following:

Permitted uses. A clear statement that AI may be used for certain functions, such as generating alternative hypotheses, brainstorming thought experiments, drafting outlines, editing prose, and assisting with code, provided that the outputs are reviewed and validated by humans.

Prohibited uses. Explicit red lines, such as using AI to fabricate data, generate fictitious references, write peer reviews, or design operational protocols in sensitive domains without proper oversight.

Responsibility. A reminder that the lab director and authors remain responsible for any claims or results, regardless of AI involvement, and that models cannot be cited as authorities.

Training. An expectation that lab members will receive basic training in AI literacy, including known failure modes, and will practice tasks with and without AI to maintain core skills.

Documentation. A recommendation that significant AI-mediated conceptual moves in a project be recorded in internal notes (for example, which prompts were used to stress-test a hypothesis), even if not all such details appear in the final paper.

Disclosure. A shared standard for when and how AI use should be mentioned in manuscripts and presentations, so that individuals are not left guessing what will be acceptable.

A charter can be one page. The point is not to anticipate every scenario, but to create a shared vocabulary and reduce ambiguity. For a younger scientist, being able to say “our lab charter encourages this” can provide some protection against ad hoc criticism, and can also serve as a gentle prompt for supervisors who have not yet thought systematically about AI.

Over time, such charters can be revisited. As norms stabilize, groups may choose to relax or tighten certain provisions. The important step is moving from unspoken expectations and private workarounds to explicit, collectively negotiated guidelines.

7.4 Navigating peer review: responding to objections about AI use

Even with careful methods descriptions and lab norms, authors will encounter reviewers and editors who are wary of AI. Responding to these objections is partly a rhetorical task and partly an opportunity to clarify and improve practice.

Objections typically fall into a few recognizable categories:

The model may have introduced errors or hallucinations. A reviewer worries that AI-suggested mechanisms or arguments are untrustworthy.

In response, authors can emphasize the screening process and, if necessary, add clarifying text. For example, in a response letter:

We agree that large language models can produce erroneous content. In this project, we used the model only to generate candidate mechanisms and questions, which were then checked against the primary literature and either developed further or discarded. We have added a sentence to the Methods section to make this screening process explicit.

A short addition to the paper might read:

All model-generated suggestions were cross-checked against published sources before inclusion.

AI use undermines originality or independence. A reviewer suggests that heavy AI involvement casts doubt on the authors' contribution.

Here, the response can reinforce the division of labor:

We view the model as a tool for expanding the space of possibilities, not as a source of conclusions. The key conceptual moves in this paper—the selection of mechanism class X, the argument linking Y to Z, and the design of the thought experiments in Section W—were developed by the authors, informed by but not dictated by model outputs. To clarify this, we have specified the limited functions of AI assistance in the Methods section.

The paper sets a bad precedent for AI use. A reviewer argues that any acknowledgment of AI in conceptual work could encourage others to overuse it irresponsibly.

In reply, authors can position their transparency as part of responsible governance:

We share the concern about irresponsible AI use. For that reason, we believe it is preferable to describe carefully constrained uses of AI, along with human oversight, rather than to omit mention of tools that are in fact widely used. Our description is intended to model a cautious, accountable approach in which AI plays a supporting role and human authors remain fully responsible for all claims.

Editors may also worry about policy compliance. If a journal has specific guidelines, authors should show how they have adhered to them and be willing to adjust wording to match house style. Often, resistance softens once it is clear that AI was used in a limited, well-understood way, not as a shortcut to avoid thinking.

In extreme cases, a reviewer may recommend rejection based primarily on the presence of AI, regardless of content. For a younger scientist, there is no perfect defense against such idiosyncratic reactions. However, consistent use of calm, method-focused language, coupled with clear evidence of rigor elsewhere in the paper, can make it easier for editors to overrule such objections.

More broadly, navigating peer review in this domain benefits from a simple stance: do not argue that AI deserves special deference, do not hide its role, and do not cede the frame that using modern tools is somehow an admission of weakness. Instead, insist quietly that what matters is the quality and accountability of the reasoning, and demonstrate through practice that centaur methods can meet the same standards of clarity, coherence, and responsiveness to critique as any other scientific approach.

8. Implications for Institutions and the Future of Scientific Work

If AI-assisted thought experiments are to move from private practice to public norm, the change cannot be left solely to individual scientists. Institutions shape what counts as legitimate training, what kinds of contributions are rewarded, which tools are supported, and how the collective memory of science is organized. Frontier AI enters that landscape as both a pressure and an opportunity.

On the one hand, institutions risk entrenching a double standard: quietly benefiting from AI-accelerated work while officially maintaining pre-AI ideals of solitary cognition. On the other hand, they can help design and stabilize centaur practice, giving younger scientists room to experiment responsibly without having to fight every battle alone. This section considers four institutional fronts: training, evaluation, infrastructure, and the longer trajectory from lone minds to networked intelligences.

8.1 Rethinking training: teaching AI-literate thought-experiment design

Training is where habits of mind are set. If students learn that using AI for conceptual work is cheating or a sign of weakness, they will either avoid it altogether or hide it, carrying a pattern of secrecy into their careers. If, instead, they are taught to integrate AI into thought-experiment design in disciplined ways, they can develop both independent judgment and tool fluency.

An AI-literate curriculum would include at least three layers.

First, foundational reasoning without AI. Students still need to wrestle with problems unaided: constructing arguments, designing thought experiments, and identifying hidden assumptions by hand. There should be assignments and exams where the use of models is explicitly forbidden, not as a nostalgic tribute to the past, but as a way of ensuring that core cognitive muscles are exercised. The ability to reason in the absence of tools remains a baseline.

Second, explicit instruction in AI as a generative instrument. Rather than leaving students to discover models on their own, courses can demonstrate how to use them to expand hypothesis space, construct counterfactual scenarios, or generate adversarial critiques. This includes teaching about failure modes: hallucination, bias, spurious structure, and overconfidence. Students can be asked to compare their own thought experiments with AI-assisted ones, noting both gains and distortions.

Third, meta-skills for orchestration and self-regulation. Students can learn when to invoke AI and when to refrain; how to log interactions; how to annotate which ideas arose from models; and how to perform post-hoc checks. They can be given tasks where the explicit goal is to design a centaur workflow: specifying roles for AI and for themselves, deciding stopping rules, and reflecting on how their own thinking changed in the process.

Institutions can support this by creating shared resources: model access in teaching labs, curated prompt libraries for conceptual work, and cross-disciplinary courses on AI in scientific reasoning. The aim is not to produce model operators who depend on prompts for every move, but scientists who understand both the power and the limits of AI as a thought-experiment partner.

8.2 Granting, hiring, and promotion in a centaur world

Evaluation systems send strong signals about what kinds of work are valued. If grant panels, hiring committees, and promotion boards implicitly penalize visible AI use in conceptual work, centaur practice will remain risky. If they adapt criteria to recognize responsible orchestration of AI, they can help redirect the culture.

One adjustment is to broaden what counts as methodological sophistication. Today, a proposal or dossier might be praised for advanced statistics, novel experimental techniques, or complex simulations. In a centaur world, careful design of AI-assisted reasoning workflows can be a comparable form of sophistication. Committees can ask: does this scientist demonstrate control over the tool, clear acceptance boundaries, and thoughtful integration of AI-generated structure with empirical and theoretical work?

Another adjustment is to avoid fetishizing “AI-free” originality. Committees should resist the temptation to treat visible AI use as a blemish or to assume that unacknowledged AI absence implies superior independence. Instead, they can look for signs of genuine intellectual ownership: the ability to explain why a particular thought experiment was framed as it was, how alternative possibilities were considered, and how the author’s own judgment shaped the final form.

Granting agencies can also set expectations in their guidelines. They might require applicants to describe any significant AI use in conceptual development, much as they require description of statistical methods or animal care protocols. This does not mean evaluating proposals on how much AI they use, but on how responsibly and transparently they do so.

Promotion criteria can be updated to recognize contributions that involve building centaur capacity: developing reusable workflows, teaching AI literacy, or creating shared conceptual tools that others can adopt. In this way, the ability to lead AI-integrated intellectual work becomes part of what it means to be a senior scientist, rather than something that must be hidden as a private convenience.

8.3 From taboo to infrastructure: specialized tools for conceptual labs

At present, many AI-assisted thought experiments take place in generic chat interfaces, attached to individual accounts, with minimal logging or structure. This reflects the taboo status of the

practice. What happens in the chat stays in the chat. From an institutional perspective, this is fragile. It makes it difficult to reproduce conceptual workflows, to audit how models were used in controversial cases, or to learn from successful centaur patterns.

Moving from taboo to infrastructure means building tools that treat conceptual work as a first-class target of support. Such tools might offer:

Versioned prompt and response histories linked to specific projects, so that the evolution of a thought experiment can be traced over time and revisited by collaborators.

Role-based multi-agent setups, where different instances of a model can be configured as skeptics, advocates, methodologists, or ethicists, with their interactions recorded as part of the scientific record.

Integration with literature databases, so that AI-suggested mechanisms or scenarios can be cross-referenced immediately with existing work, reducing the risk of hallucinated novelty.

Local or institutional deployments that meet data governance requirements, allowing sensitive conceptual work to be done without exporting proprietary or pre-publication content to external servers.

User interfaces designed for scientists rather than casual chat, emphasizing experiment design, logging, and comparison over free-form conversation.

Institutions that invest in such infrastructure signal that AI-assisted thought experiments are not an embarrassing side-channel but a legitimate part of the research environment. They also gain leverage for governance: logging and structure make it easier to enforce red lines in dual-use contexts and to investigate problematic uses if they arise.

8.4 Long-term trajectories: from lone minds to networked intelligences

Looking beyond immediate policies and tools, AI-assisted thought experiments point toward a broader shift in how scientific work is organized. The traditional picture centers on the lone mind or small team, interacting with instruments and data but fundamentally self-contained. Frontier AI, especially when deployed as part of connected infrastructures, nudges us toward a model of distributed, networked intelligences.

In such a model, individual scientists, AI systems, and institutional memories form a composite cognitive field. Ideas do not originate in a single mind so much as emerge from patterns of interaction: human–human, human–machine, and machine–machine. Thought experiments may be designed, critiqued, and refined across multiple sites and timescales, with models carrying forward certain structural motifs while humans provide anchoring in empirical reality and moral judgment.

This trajectory has both promise and danger. On the promising side, it could accelerate the exploration of complex, multi-level problems that no single mind or lab could handle alone. Networked centaur systems could coordinate thought experiments across disciplines, discovering structural rhymes between, for example, climate dynamics and cellular regulation, or between legal doctrine and algorithmic governance. The imaginative capacity of science would increase, not just in volume but in connectivity.

On the dangerous side, there is a risk of homogenization and over-reliance. If many nodes in the network depend on similar models, training data, or interface designs, the space of thought experiments may narrow even as it appears to expand. Certain framings could become dominant not because they are true, but because they are easy for models to produce and for institutions to recognize. Human judgment could atrophy if too much trust is placed in the patterns of a network whose biases are hard to see from within.

Institutions have some choice in how this trajectory unfolds. They can foster diversity in models and methods, support pluralism in conceptual frameworks, and encourage cross-examination between different centaur systems. They can design governance structures that keep human responsibility and moral reflection at the center, even as much of the combinatorial work of idea generation is automated.

For younger scientists, the long-term message is sobering and hopeful. The era of the isolated genius is fading, but that does not mean individuals are irrelevant. Instead, the distinctive contribution of a scientist may lie increasingly in how they navigate and shape a field of networked intelligences: how they choose which tools to trust, which thought experiments to pursue, which institutional structures to build or challenge. Centaur practice is not a temporary stage on the way to being replaced; it may be the enduring form of human agency in a scientifically saturated, AI-infused world.

9. Conclusion

Frontier AI has already changed what it means to think scientifically, whether or not institutions have caught up. A single researcher, with access to a capable model, can now explore vast spaces of hypotheses, counterfactuals, and arguments in a fraction of the time that such exploration once required. In the safe regime of thought experiments—where outputs are ideas subject to human acceptance rather than instructions for direct action—this capacity is less a threat than an invitation. It offers to expand the reach of imagination under human responsibility.

Yet the full potential of this shift is not being realized. Instead, scientific culture is caught in an uncomfortable in-between. Many researchers, especially younger ones, rely heavily on AI in

private, while public narratives and official norms still pretend that core conceptual work happens in essentially the same way it did before. The result is a growing gap between practice and story, and a set of gatekeeping behaviors that protect existing hierarchies more than they protect the integrity of inquiry.

This conclusion gathers the main threads of the argument. It recaps what is already possible and what is being blocked, outlines the costs of continued hypocrisy and under-use, sketches a constructive path toward normalizing AI-assisted thought experiments, and ends with reflections on the kinds of courage and responsibility that centaur practice requires.

9.1 What is already possible, and what is being blocked

Within the boundaries described in this paper, frontier AI is already a capable thought-experiment instrument. It can instantiate multiple perspectives on a problem, enumerate alternative explanations and failure modes, generate counterfactual scenarios, and act as an adversarial critic. It can help structure complex arguments, suggest unusual analogies, and expose hidden assumptions. In principle, it can democratize access to a level of conceptual scaffolding that once depended on proximity to a small number of unusually broad and practiced minds.

Importantly, all of this is possible without giving the model any direct control over experiments, data collection, or policy decisions. AI remains on the side of proposal and critique, while humans retain control over acceptance and action. In this regime, the main technical risks—hallucination, bias, spurious coherence—are real but manageable with existing scientific disciplines: cross-checking, skepticism, and explicit separation of speculation from result.

What is being blocked is not the tool's capacity but its legitimate visibility. Gatekeeping behaviors, rooted in cognitive monopoly protection and status preservation, constrain who can safely admit to using AI in conceptual work and how. Younger scientists are pressured, implicitly or explicitly, to confine AI to low-status tasks such as editing and boilerplate code, even as they rely on it for more substantive thinking. Senior scientists may quietly use AI for thought experiments while publicly reinforcing norms that treat such use as suspect.

This blocking does not prevent AI-assisted thought experiments from happening. It merely forces them into the shadows, where they cannot be collectively refined, critically examined, or used to improve methods. The very place where AI could help most—the public, accountable space of reasoning—is where it is most constrained by unspoken taboos.

9.2 The costs of prolonged hypocrisy and under-use

Maintaining a double life around AI use is not cost-free. It carries cumulative harms for individuals, institutions, and the scientific enterprise as a whole.

One cost is distortion of the scientific record. When conceptual workflows that rely heavily on AI are reported as if they were purely human, the literature becomes a kind of historical fiction. Future readers cannot tell which arguments emerged from extended human debate and which were shaped by model-generated structures. Methodologists cannot study how best to integrate AI if that integration remains invisible.

Another cost is inequity. Senior researchers, protected by reputation and tenure, can use AI aggressively in private with little risk, while junior researchers face greater scrutiny and potential punishment for visible use. This asymmetry multiplies existing power imbalances. Those already in positions of authority can quietly accelerate their work, while those trying to climb the ladder must either handicap themselves or participate in a system of concealment that could later be used against them.

A third cost is lost training value. When AI use is taboo, students and early-career scientists do not receive explicit instruction in how to use it well. They are left to discover models on their own, without guidance on failure modes, orchestration, or appropriate boundaries. As a result, some become over-dependent and shallow in their use, while others miss opportunities to expand their thinking responsibly. An unacknowledged tool cannot be consciously mastered.

There is also an ethical cost. Hypocrisy erodes trust. When scientific culture encourages people to use tools privately while pretending not to publicly, it normalizes a kind of double bookkeeping that can spill over into other domains. If it is acceptable to omit AI from methods because “everyone does it,” it may become easier to rationalize other omissions or ambiguities. Transparency about AI is not just a technical issue; it is part of the broader fabric of honesty in science.

Finally, under-use in the safe regime slows progress on genuinely hard problems. Many of the most pressing scientific and social challenges are structurally complex and cross-disciplinary. AI-assisted thought experiments offer a way to explore such complexity more thoroughly, surfacing intersections and edge cases that single minds might miss. Treating this capacity as marginal or suspect leaves potential insight on the table.

9.3 A constructive path: normalizing AI-assisted thought experiments

The alternative to hypocrisy and under-use is not naive enthusiasm or surrender of human responsibility. It is a deliberate normalization of AI-assisted thought experiments within a centaur framework. That normalization can proceed along several practical lines.

At the level of individual practice, scientists can adopt clear internal norms. Use AI for proposal, not proof. Maintain a sharp acceptance boundary where human judgment must be exercised and documented. Treat model outputs as objects of scrutiny, not as sources of authority. Keep

private notes of which ideas and structures arose from AI interactions, even if only a subset of that information appears in the final publication.

In written work, authors can begin to use neutral, method-like descriptions of AI use. Short methods subsections that specify what the model was used for, what it was not used for, and how its outputs were screened can make AI assistance visible without turning it into a spectacle. Over time, such descriptions can converge into field-specific patterns that feel as routine as descriptions of statistical procedures or simulation protocols.

At the group level, labs can write simple charters for AI use: what is permitted, what is prohibited, and how responsibility is understood. These charters can be living documents, updated as experience accumulates. They can also serve as training tools, giving students a framework for thinking about AI that goes beyond either fear or uncritical adoption.

Institutions can support these shifts by updating training, evaluation, and infrastructure. Courses can teach AI-literate thought-experiment design, not just AI for coding or data analysis. Grant and hiring guidelines can ask for concise descriptions of AI use without penalizing responsible disclosure. Shared tools can be developed for logging and structuring conceptual interactions with models, treating them as part of the research apparatus rather than as private, unrecorded conversations.

None of this eliminates the need for vigilance around misuse. Certain domains will continue to require strict constraints on AI involvement, especially where the line between hypothetical and operational content is thin. But within the safe regime, the guiding principle can be: use the tool fully, describe it honestly, and keep the human accountable.

9.4 Final reflections on courage, responsibility, and shared intellect

What this paper ultimately calls for is not a technical adjustment so much as a shift in intellectual culture. It asks scientists to acknowledge that their minds are no longer the only engines of structured imagination in their work, and to respond to that fact not with denial but with a new form of courage and responsibility.

Courage, because stepping into centaur practice visibly means running ahead of some current norms. It may invite suspicion from those invested in older images of thinking. It requires admitting, in public, that one's ideas were shaped in part by a machine, and trusting that the value of the work lies in how one has judged and integrated those influences, not in maintaining a fiction of solitary genius.

Responsibility, because AI does not absolve anyone of the duty to think carefully. If anything, it increases the need for discipline. The more powerful the generator of possibilities, the more important it is to curate them wisely, to resist the pull of plausible but hollow structure, and to

ensure that thought experiments remain grounded in empirical knowledge and ethical reflection. The human role at the acceptance boundary becomes both more demanding and more central.

Finally, there is an invitation to embrace a new sense of shared intellect. Scientific thinking has always been a collective enterprise, distributed across people, institutions, and time. Frontier AI adds a new kind of participant to that collective: not a person, not an author, but a pervasive patterning force in how possibilities are generated and combined. The question is not whether this presence will shape the future of thought experiments, but whether it will do so in ways that are visible, governed, and aligned with the values of inquiry.

To move forward, scientific communities do not need to choose between human minds and machines. They need to design and inhabit configurations in which humans, aided by machines, can think more widely without thinking less deeply. Normalizing AI-assisted thought experiments within a centaur framework is one step in that direction: a way of expanding the space of what can be imagined while preserving, and even sharpening, the human responsibility for what is accepted as true.

10. References

1. Luo, Z., Yang, Z., Xu, Z., Yang, W., & Du, X. (2025). LLM4SR: A survey on large language models for scientific research. *ACM Computing Surveys* (in press). arXiv
2. Zhang, Y., Chen, X., Jin, B., Wang, S., Ji, S., Wang, W., & Han, J. (2024). A comprehensive survey of scientific large language models and their applications in scientific discovery. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2024). ACL Anthology
3. Ahn, S., Liao, W.-H., Kim, A., et al. (2024). The transformative impact of large language models on medical writing and publishing. [Journal article]. PMC
4. McCracken, C. (2025). Is generative artificial intelligence changing research practices? Coalition for Networked Information / Ithaka S+R briefing, 24 January 2025. CNI: Coalition for Networked Information
5. Mohammadi, E., et al. (2025). Is generative AI reshaping academic practices worldwide? Evidence from faculty and student surveys. *Scientometrics* (in press). ScienceDirect
6. Conroy, G. (2023). How ChatGPT and other AI tools could disrupt scientific publishing. *Nature*, 622, 234–236. doi:10.1038/d41586-023-03144-w. Nature
7. Dang, A.-H., Tran, V., & Nguyen, L.-M. (2025). Survey and analysis of hallucinations in large language models: Attribution to prompting strategies or model behavior. *Frontiers in Artificial Intelligence*, 8, 1622292. doi:10.3389/frai.2025.1622292. PMC
8. Cossio, M. (2025). A comprehensive taxonomy of hallucinations in large language models. arXiv preprint arXiv:2501.00999. arXiv
9. Liang, T., He, Z., Jiao, W., et al. (2024). Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of EMNLP 2024*. ACL Anthology
10. Liu, Y., Cao, J., Li, Z., He, R., & Tan, T. (2025). Breaking mental set to improve reasoning through diverse multi-agent debate. In *Proceedings of ICLR 2025*. OpenReview
11. Yang, W., Chen, B., Ren, X., Yang, Y., et al. (2025). A survey on integrating large language models and knowledge bases. arXiv preprint. ScienceDirect
12. Zhang, X., Chen, Y., Li, J., et al. (2024). Scientific large language models: Current progress and future directions. [Preprint]. ACL Anthology
13. Yin, W., Chen, B., Luo, X., et al. (2024). A survey on multimodal large language models. *Machine Intelligence Research*, 21(3), 1–30. OUP Academic

14. Vasconcelos, S. (2025). Generative AI and research integrity: Emerging challenges for scholarly publishing. *Accountability in Research* (advance online publication). PMC

15. Frangou, S. (2025). Artificial intelligence in scientific writing and publishing: A call for critical engagement. *Biological Psychiatry* (commentary). PMC

16. American Chemical Society (ACS Publications). (2024). Artificial intelligence (AI) best practices and policies at ACS Publications. Policy statement. researcher-resources.acs.org

17. EDP Sciences. (2025). EDP Sciences updates its policy on AI and ethics. Publisher policy update. edpsciences.org

18. Wolters Kluwer – Lippincott Journals. (2025). Ethical best practices in scholarly publishing: Artificial intelligence authoring tools and authorship policy. In *Ethical Best Practices in Scholarly Publishing*. Wolters Kluwer

19. Committee on Publication Ethics (COPE). (2023). Artificial intelligence (AI) in decision making and scholarly communication: Discussion documents and position statements. Wolters Kluwer

20. Nature Editorial and News Staff. (2025). AI is saving time and money in research — but at what cost? *Nature* (news feature). Nature+1

The author has published over 4,600 AI-assisted papers at: https://www.researchgate.net/profile/Douglas-Youvan/research . Google Scholar has indexed these preprints, and if you want a particular reference, the AI mode of a Google search (Gemini) has efficiently catalogued these preprints.